L'information biographique : modélisation, extraction et organisation en base de connaissances

Laurent Kevers

Université catholique de Louvain – CENTAL
laurent.kevers@uclouvain.be

Résumé

L'extraction et la valorisation de données biographiques contenues dans les dépêches de presse est un processus complexe. Pour l'appréhender correctement, une définition complète, précise et fonctionnelle de cette information est nécessaire. Or, la difficulté que l'on rencontre lors de l'analyse préalable de la tâche d'extraction réside dans l'absence d'une telle définition. Nous proposons ici des conventions dans le but d'en développer une. Le principal concept utilisé pour son expression est la structuration de l'information sous forme de triplets {sujet, relation, objet}. Le début de définition ainsi construit est exploité lors de l'étape d'extraction d'informations par transducteurs à états finis. Il permet également de suggérer une solution d'implémentation pour l'organisation des données extraites en base de connaissances.

Mots-clés : information biographique, modélisation, extraction d'information, transducteur à états finis, entité nommée, relation, base de connaissances.

Abstract

Extraction and valorization of biographical information from news wires is a complex task. In order to handle it correctly, it is necessary to have a complete, accurate and functional definition. The preliminary analysis of the extraction task reveals the lack of such a definition. This article proposes some conventions to develop it. Information modelling as triples {subject, relation, object} is the main concept used at this level. This incomplete definition can be used during the information extraction step. It also allows to suggest some implementation solutions for data organisation as a knowledge base.

Keywords: biographical information, modelling, information extraction, finite state transducers, named entities, relation, knowledge base.

1. Introduction

Les textes journalistiques se caractérisent souvent par une proportion élevée de noms propres. Selon (Fourour, 2004), les anthroponymes en constituent la catégorie la plus importante (de 50 % à 70 % des formes en fonction des types de journaux). Notre travail sur l'information biographique repose sur ces constats. Il se base sur un large corpus de dépêches de presse fournies par l'agence de presse belge Belga et s'articulera en quatre points :

- la définition même de l'information biographique ;
- le formalisme à utiliser pour exprimer les faits ;
- l'apport d'une définition claire pour la phase d'extraction ;
- la manière de conserver les informations extraites.

Lorsqu'il s'agit d'extraire, de modéliser et de structurer des données biographiques, on se heurte rapidement à une question de taille : « Qu'est-ce que réellement l'information

biographique ? » Tout le monde sait intuitivement ce qu'est une biographie, mais il est assez malaisé d'en définir précisément le contenu. Plusieurs ressources disponibles sur Internet (Davis, 2004 ; Kanzaki, 2003) proposent des nomenclatures, mais celles-ci sont forcément incomplètes. Le niveau de généricité est également peu élevé car ces documents sont construits dans une optique d'implémentation plutôt que de modélisation. Le flou en ce qui concerne cette définition est assez gênant, que ce soit pour le développement de la phase d'extraction ou pour l'organisation du stockage des informations. Il est en effet impossible de travailler efficacement si l'on ne connaît pas avec précision l'objet de l'étude. La clarification du domaine d'application profite à l'ensemble du processus, depuis l'extraction d'information jusqu'à la phase d'accumulation des données.

La première partie sera donc consacrée à une approche intuitive de l'information biographique et à la définition des concepts sous-jacents à ces intuitions. À partir de ces définitions, une nomenclature de l'information biographique peut être construite. La deuxième section s'attaquera à cette tâche, sans avoir la prétention de l'exhaustivité. La troisième partie montre l'apport de l'analyse des évènements biographiques lors du développement de ressources d'extraction. Enfin, dans une dernière section, toujours sur la base des concepts mis en évidence dans la deuxième partie, le choix d'une méthode de stockage des données sera suggéré.

2. Approche intuitive et définition

Une biographie est définie par le *Petit Robert* (2004) comme étant un « écrit qui a pour objet l'histoire d'une vie particulière ». Autrement dit, il s'agit des évènements survenant dans la vie des personnes. Ces évènements sont en relation avec différents éléments de leur vie quotidienne. Ils font intervenir des personnes ou des organisations, sont caractérisés par des dates et des lieux, impliquent divers « objets » plus ou moins abstraits avec lesquels ils sont amenés à interagir.

Cette première perception mérite une formalisation plus précise. Nous appellerons « entité » les éléments intervenant dans les données biographiques (personnes, organisations, lieux, dates, etc.). Ce concept est défini dans (Chinchor, 1998), mais l'interprétation utilisée ici sera moins stricte. Les types d'entités seront plus nombreux et pourront inclure certains éléments exclus dans cette définition. Nous appellerons « évènement » toute action faisant intervenir ces entités. L'analyse d'un évènement de la vie réelle permet de décomposer celui-ci en plusieurs « relations ». Une relation lie deux entités quelconques. L'une d'elle joue le rôle de « sujet », alors que la seconde constitue l'« objet » de la relation. Un évènement est donc formalisable sous la forme d'un ensemble de triplets {sujet, relation, objet}. Cette approche, que l'on peut retrouver dans les travaux portant sur le Web sémantique (Charlet *et al.*, 2002), a également été suggérée dans le cadre de travaux en extraction d'informations (Bouhafs, 2004 ; Le Priol, 2001). Une entité qui joue un rôle d'objet dans une relation peut être le sujet d'une autre relation, et inversement. L'entité centrale pour l'information biographique est bien entendu la personne. Elle sera par conséquent souvent utilisée en tant que sujet des relations.

La décomposition des faits biographiques en relations ne veut pas dire que l'on se débarrasse complètement de la notion d'évènement. En effet, les relations découlant d'un évènement doivent toujours être interprétées conjointement. Prises individuellement ou en combinaison avec des relations issues d'autres évènements, elles mènent à des interprétations incomplètes ou erronées.

À l'aide de ces concepts, il est possible de définir ce qu'est l'information biographique. La partie suivante est consacrée au recensement de quelques évènements. Cette liste permet de

prendre la mesure du problème et d'établir concrètement un format de spécification des évènements, des entités et des relations.

3. Information biographique, évènements et relations entre entités

Pour chaque évènement, la liste des relations qui en découle est développée. Une cardinalité indique la fréquence d'apparition des évènements par rapport à une personne. Si nécessaire, des contraintes supplémentaires sont introduites. Toute relation implique l'existence de son inverse. Pour l'évènement « naissance », la relation *X a pour parent Y* implique que *Y est parent de X*. Il n'est par conséquent pas nécessaire de prévoir cette dernière dans la définition. La spécification se situe à un niveau conceptuel et non linguistique. Elle définit ce dont l'information est composée, mais pas la manière dont celle-ci sera effectivement exprimée dans les textes.

3.1. Informations personnelles

Naissance [1-1]

relation	***sujet***	***objet***
X se nomme N	X: personne	N: nom
X se prénomme P	X: personne	P: prénom
X est de sexe S	X: personne	S: sexe
X est né le D	X: personne	D: date
X est né à L	X: personne	L: lieu
X a pour parent Y	X: personne	Y: personne
X est de nationalité N	X: personne	N: nationalité
X est identifié par R	X: personne	R: n° de registre national

Décès [0-1]

relation	***sujet***	***objet***
X est décédé le D	X: personne	D: date
X est décédé à L	X: personne	L: lieu
X est décédé de C	X: personne	C: cause de décès

Mariage [0-N]

relation	***sujet***	***objet***
X s'est marié le D	X: personne	D: date
X s'est marié à L	X: personne	L: lieu
X est marié avec Y	X: personne	Y: personne

Divorce [0-N]

relation	***sujet***	***objet***
X a divorcé le D	X: personne	D: date
X a divorcé à L	X: personne	L: lieu
X est divorcé de Y	X: personne	Y: personne

Concubinage [0-N]

relation	***sujet***	***objet***
X est concubin de Y	X: personne	Y: personne
X est concubin depuis D	X: personne	D: date
X est concubin jusque D'	X: personne	D': date

La cardinalité de *divorce* est liée à la cardinalité de *mariage*. Pour *divorce*, la borne inférieure est toujours égale à zéro ou au nombre de mariages moins un si le nombre de mariages est supérieur ou égal à un. La borne supérieure est toujours égale au nombre de mariages.

3.2. Informations relatives à la formation

Obtention d'un diplôme [0-N]

relation	***sujet***	***objet***
X a obtenu le niveau d'étude Y	X: personne	Y: diplôme
Y a été obtenu le D	Y: diplôme	D: date
Y est un diplôme délivré par O	Y: diplôme	O: organisation

3.3. Informations professionnelles

Occupation une fonction, d'un poste [0-N]

relation	*sujet*	*objet*
X exerce le profession de M	X: personne	M: profession
X fait partie de O	X: personne	O: organisation
X est engagé le D	X: personne	D: date
X est remercié le D'	X: personne	D': date

Création d'une entreprise [0-N]

relation	*sujet*	*objet*
X a créé O	X: personne	O: organisation
O a été créée le D	O: organisation	D: date

Cession d'une entreprise [0-N]

relation	*sujet*	*objet*
O a été vendu le D	O: organisation	D: date
X a vendu O	X: personne ou organisation	O: organisation
O a été vendu à Y	O: organisation	Y: personne ou organisation

3.4. Informations relatives à des récompenses

Obtention d'une distinction, récompense [0-N]

relation	*sujet*	*objet*
X a reçu la distinction R	X: personne	R: récompense, prix, distinction
R est attribué par O	R: récompense, prix, distinction	O: organisation
R a été attribué le D	R: récompense, prix, distinction	D: date

Victoire lors d'une compétition, d'un concours [0-N]

relation	*sujet*	*objet*
X a remporté C	X: personne ou organisation	C: compétition, concours
C a lieu à L	C: compétition, concours	L: lieu
C a lieu le D	C: compétition, concours	D: date

3.5. Informations juridiques

Dépôt d'une plainte [0-N]

relation	*sujet*	*objet*
X a déposé plainte pour A	X: personne ou organisation	A: motif d'accusation
X a déposé plainte contre Y	X: personne ou organisation	Y: personne ou organisation
X a déposé plainte auprès de O	X: personne ou organisation	O: organisation
X a déposé plainte le D	X: personne ou organisation	D: date

Arrestation [0-N]

relation	*sujet*	*objet*
X a été arrêté pour A	X: personne	A: motif d'arrestation
X a été arrêté par Y	X: personne	Y: personne ou organisation
X a été arrêté à L	X: personne	L: lieu
Y a été arrêté le D	X: personne	D: date

Inculpation [0-N]

relation	*sujet*	*objet*
X a été inculpé pour I	X: personne ou organisation	I: motif d'inculpation
X a été inculpé par O	X: personne ou organisation	O: personne ou organisation
Y a été inculpé le D	X: personne ou organisation	D: date

Condamnation [0-N]

relation	*sujet*	*objet*
X a été condamné pour C	X: personne ou organisation	C: motif de condamnation
X a été condamné par O	X: personne ou organisation	O: personne ou organisation
X a été condamné à P	X: personne ou organisation	P: peine
Y a été condamné le D	X: personne ou organisation	D: date

4. Développement des ressources d'extraction

4.1. Processus général

Le processus général d'extraction d'informations biographiques présenté ici constitue une première approche qui doit certainement être améliorée. Divers travaux peuvent être consultés à cet effet. Par exemple, (Grishman, 1997) expose les principes généraux des systèmes d'extraction d'informations et (Fourour, 2004) propose un état de l'art ainsi qu'une implémentation d'un système mixte à base de règles et d'apprentissage. Les premiers résultats obtenus permettent d'avoir une idée plus précise de l'ensemble des traitements à effectuer dans le cadre de l'extraction et de l'exploitation d'informations biographiques.

La technique choisie, et présentée par exemple par (Poibeau *et al.*, 1999), fait appel à des transducteurs à états finis. Ceux-ci permettent une description et une annotation des motifs que l'on désire retrouver. Il s'agit d'une analyse locale dont l'action se situe au niveau subphrastique. Plusieurs traitements sont effectués au préalable sur les textes : découpe en tokens et en phrases, application de dictionnaires. Ces manipulations sont effectués à l'aide d'Unitex (Paumier, 2004).

Le processus général se décompose en plusieurs passes, chacune correspond à un niveau de complexité des éléments recherchés. Les premières passes consistent en l'application de graphes qui exploitent les informations lexicales ainsi que des indices internes (la structure des entités) et externes (le contexte des entités) tels que ceux présentés dans (McDonald, 1996). Le but est de retrouver et d'annoter des entités de base, souvent appelées entités nommées, telles que les personnes, les organisations, les lieux... Les passes suivantes s'appuient sur cette première analyse pour rechercher des informations disposées de manière de plus en plus complexe dans les phrases. Ce mécanisme s'inspire de l'approche par cascade d'automates décrite par (Friburger *et al.*, 2004). Ce processus d'applications successives de transducteurs permet de simplifier l'expression des règles de plus haut niveau, c'est-à-dire celles passant à la fin. Divers traitements peuvent être introduits entre certaines passes afin d'améliorer la couverture ou la précision de l'analyse : recherche des variations graphiques et de coréférences des entités, résolution des anaphores pronominales, etc. Il s'agit de tâches complexes qui n'ont pu être abordées que de manière très superficielle pour l'instant dans le cadre de ce travail.

Une fois toutes les ressources d'extraction exploitées, le format de sortie doit d'être suffisamment général afin d'être exploitable par le plus grand nombre d'applications. Cette exigence est remplie par un fichier XML qui reprend les éléments annotés.

4.2. Extraction des entités de base

Le travail mené sur des textes journalistiques en français par (Fairon *et al.*, 2003) et en anglais par (Mallchok, 2004) a prouvé l'adéquation des transducteurs à états finis pour le repérage des entités de base. Le développement de quelques dizaines de graphes couplé à l'utilisation de ressources lexicales spécialisées permet d'obtenir une analyse telle que celle reprise ci-dessous :

> ***{BOGOTA,.PLACE+TOWN}*** *07/04 (****{AFP,.ORG}****) = Dix-sept militaires colombiens ont été tués* ***{mercredi,.DATE}*** *lors d'une embuscade de rebelles des* ***{Forces armées révolutionnaires de Colombie,.ORG}*** *(****{FARC,.ORG}****, guérilla marxiste) dans le nord de la* ***{Colombie,.PLACE+COUNTRY}*** *, a annoncé à l'****{AFP,.ORG}*** *le commandant de l'****{armée colombienne,.ORG}****, le général* ***{Reinaldo Castellanos,.PERSON}****.{S}*

Figure 1. Annotation des entités de base

Le format d’annotation adopte les conventions utilisées pour les entrées de dictionnaire DELA d’Unitex.

4.3. Extraction des informations contenues dans les appositions

Les contextes immédiats des noms de personnes sont riches en informations biographiques. On y retrouve couramment, en apposition, des données telles que l’âge, la profession, la nationalité, un titre ou une tendance politique. Il est possible de regrouper tout ces éléments en un seul groupe sans modifier la structure globale de la phrase. À l’intérieur de cette entité complexe, on conserve l’identification des différents éléments reconnus. À partir de l’exemple obtenu suite à la première passe, on pourra ainsi obtenir ce texte dans un deuxième temps :

{BOGOTA,.PLACE+TOWN} 07/04 ({AFP,.ORG}) = Dix-sept militaires colombiens ont été tués {mercredi,.DATE} lors d’une embuscade de rebelles des {Forces armées révolutionnaires de Colombie,.ORG} ({FARC,.ORG}, guérilla marxiste) dans le nord de la {Colombie,.PLACE+COUNTRY} , a annoncé à l’{AFP,.ORG} le ***{[[commandant#FCT] [armée colombienne#ORG]#PRO][le général#TITLE][Reinaldo Castellanos#NAME],.PERSON}*** *.{S}*

Figure 2. Annotation des informations en apposition

Certains regroupements ont déjà été effectués. Ce résultat, intéressant en soi, facilitera l’analyse de surface nécessaire lors de la suite du processus d’extraction.

4.4. Extraction d’informations par type d’évènement

Au delà de l’exploitation des informations contenues dans les appositions, la tâche à accomplir ensuite devient plus complexe. Non seulement l’information à extraire sera éparpillée dans toute la phrase, voire dans un paragraphe, mais elle sera aussi exprimée de façon plus variée. C’est à ce moment que le travail de définition de l’information biographique mené en amont peut réellement aider à structurer et à orienter l’extraction. Sur la base d’un type d’évènement particulier, comparable à un *scénario* selon la terminologie MUC (Grishman, 1997), il est possible de dériver des « patrons d’extraction ». La réflexion peut se faire en deux temps. Premièrement, on sélectionne un évènement et on recherche les termes dénotant cette sémantique. Ensuite, l’étude de ces derniers permet d’aboutir à un ou plusieurs motifs d’extraction. Le but étant ici de reconnaître les contextes porteurs d’informations.

Événement : mariage
Termes possibles : union, unir, épouser, époux, épouse, marier, mari, femme
Motif d’extraction (simplifié) :

Exemple de résultat :

{[prince#TITLE][Ernst-August de Hanovre#NAME] **[[[princesse#TITLE][Caroline de Monaco#NAME] #PERSON] #MARIAGE]** ,.PERSON}

Étant donné la complexité de cette étape, il n’est pas évident que celle-ci pourra être menée à bien au seul moyen de la technique d’extraction présentée ci-dessus. En effet, on constate qu’au moins l’information est localisée à un seul endroit de la phrase au moins l’approche par transducteurs, ou grammaires locales, semble pertinente. Pour les développements futurs, une

autre approche encore à définir devra probablement être étudiée afin d'évaluer les gains de performance possibles, tant au niveau de la construction même des ressources que de la qualité de l'extraction.

5. Vers une solution d'implémentation pour le stockage

5.1. Précision des concepts de relation et d'entité

L'analyse partielle de la nature des informations biographiques réalisée permet de mettre en évidence quelques caractéristiques qui peuvent nous guider dans les choix d'implémentation d'un système de stockage de ces données. Ce système sera nommé de manière générique « base de connaissances ».

D'une manière générale, on constate que les exemples d'informations biographiques mentionnés dans les sections précédentes sont effectivement bien exprimables sous la forme d'une ou plusieurs relations entre des sujets et des objets. Il est cependant nécessaire de fournir des contraintes d'intégrité afin de garantir la cohérence des données. Ces contraintes doivent déterminer quelles sont les relations utilisables pour décrire l'information biographique et entre quelles entités elles peuvent survenir. L'analyse intuitive donne déjà une assez bonne idée des relations et des contraintes à exprimer, mais plusieurs points méritent d'être spécifiés plus précisément.

En plus des relations porteuses d'informations biographiques directement inspirées de la liste donnée plus haut, nous souhaiterons également stocker un ensemble de relations dont la fonction est de donner de l'information sur l'information (méta-données). Les relations peuvent donc appartenir à deux classes différentes : les relations informationnelles et les méta-relations. On dispose par exemple de relations permettant d'indiquer un poids (ou indice de confiance) attribué à une donnée, d'indiquer la source et la date relatives à l'origine de l'information, de fournir la langue dans laquelle elle a été exprimée, de relier deux informations synonymes, etc. Selon les besoins particuliers, on pourra encore ajouter différentes méta-relations.

Les entités peuvent également être réparties en différents types. À l'instar des relations, il existe des entités informationnelles et des méta-entités. En pratique, il n'y a pas une grande différence entre ces deux types d'entités si ce n'est qu'une méta-entité est toujours associée à une méta-relation. Les natures des entités informationnelles peuvent être dérivées de l'analyse de la première partie et celles des méta-entités sont reprises dans la table ci-dessous.

méta-relation	***nature de X***
I a un indice de confiance de X	Un entier ou un réel
I est en langue X	Un code représentant une langue
I a été ajouté le X	Une date
I provient de la source X	Une valeur désignant un document
I est un synonyme de X	Une information reprise dans la base de données

Figure 3. Méta-relations. Soit I une information contenue dans la base de connaissances

La difficulté qui apparaît avec la décomposition en relations telle que présentée au paragraphe 3 est la conservation de la cohérence et de l'intégrité des données. Prenons l'exemple d'un

mariage entre monsieur Smith et mademoiselle Dupond, qui a lieu le 17 juillet 2007 à Bruxelles. On peut décomposer cet évènement avec les relations suivantes[1] :

M. Smith *s'est marié à* Bruxelles	Mlle Dupond *s'est marié à* Bruxelles
M. Smith *s'est marié le* 17 juillet 2007	Mlle Dupond *s'est marié le* 17 juillet 2007
M. Smith *s'est marié avec* Mlle. Dupond	Mlle Dupond *s'est marié avec* M. Smith

Le problème de cette décomposition, c'est que Mlle Dupond et M. Smith peuvent être impliqués dans plusieurs mariages au cours de leur vie. Il sera alors impossible de savoir à quel mariage correspond quelle date et quel lieu. Pour modéliser l'information de manière correcte, il faut énoncer les relations de la manière suivante :

M. Smith *s'est marié avec* Mlle Dupond Mlle Dupond *s'est marié avec* M. Smith
Le mariage de M. Smith et de Mlle Dupond *a eu lieu le* 17 juillet 2007
Le mariage de M. Smith et de Mlle Dupond *a eu lieu à* Bruxelles

Dans cet exemple, « le mariage de M. Smith et de Mlle Dupond » est une entité composée à partir d'une relation (*s'est marié avec*) entre deux entités (deux personnes). Cela nous amène à élargir le concept de relation en admettant qu'une première relation peut jouer le rôle de sujet dans une seconde.

5.2. Évaluation de la pertinence du modèle de données en vue d'une implémentation

Un des avantages de la structuration de l'information sous forme de relations est qu'elle présente un haut degré de généricité. La modélisation des données, quelles qu'elles soient, à l'aide d'un triplet permet de placer la sémantique uniquement au niveau des données et non dans la structure de celles-ci. Les bases de données relationnelles sont souvent conçues en définissant des types d'entités et de relations, plus tard traduits en tables. Dans ce cas, la structure des tables contient une partie de l'information ! L'utilisation de ce genre de système n'est pas en soi un problème lorsqu'on connaît bien le domaine d'application. En ce qui concerne les informations biographiques, tout porte à croire qu'il sera pratiquement impossible d'arrêter une structure complète et définitive avant toute implémentation. L'adoption d'une structuration sous forme de triplets devrait nous apporter la souplesse nécessaire à l'adaptation continue de la modélisation du domaine d'application. En effet, dans un système de base de données relationnelle classique, l'ajout de nouveaux attributs à une entité devrait se traduire par une altération de la structure d'une ou plusieurs tables, alors que le système envisagé ne demandera que l'ajout d'un triplet.

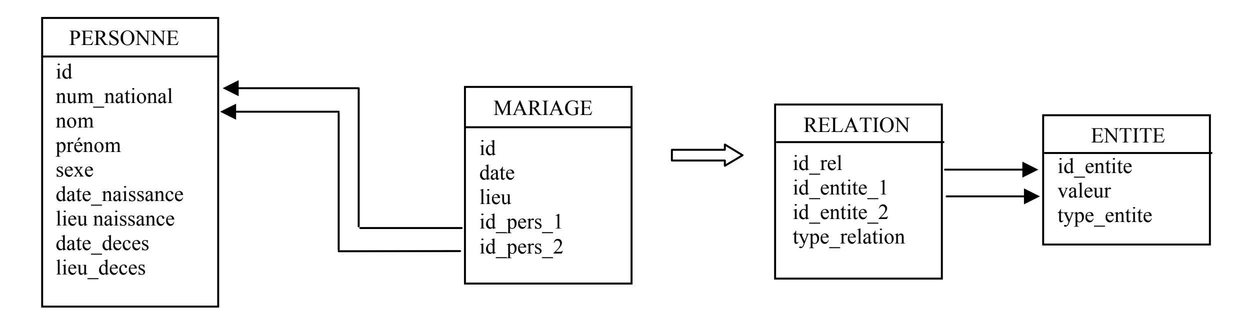


Figure 4. Base de données relationnelle classique et base de données « générique »

Si d'un point de vue informatique, une forte formalisation des données est généralement souhaitable, elle se fait parfois quelque peu au détriment de la richesse de celles-ci.

[1] *S'est marié à*, *s'est marié le* et *s'est marié avec* sont des noms de relations. Ces noms ne s'accordent pas selon le genre ou le nombre de l'entité qui joue le rôle de sujet.

L'organisation des données autour d'un nombre restreint de relations bien définies implique une certaine perte au niveau de la formulation originale (par rapport au texte de départ, en langage naturel). Lors de l'exploitation des informations, il faut être conscient qu'un certain nombre d'applications nécessiteront éventuellement une reformulation vers le langage naturel, ce qui n'est pas une tâche des plus triviale.

Le fait que les relations soient codées comme des valeurs dans une structure générique impose une grande rigueur en ce qui concerne l'encodage de celles-ci. Toute erreur à cet endroit rendrait une partie de l'information inutilisable. Des mécanismes de contrôle (vérification de contraintes d'intégrité) doivent être mis en place pour s'assurer de l'emploi correct d'une relation.

Une conséquence de la généricité du système est que toutes les informations seront stockées quasiment au même endroit. Cela va nous amener à créer et manipuler quelques objets (tables) très volumineux. Comme toujours dans ce cas, la question de la performance se pose.

5.3. Pistes retenues pour une implémentation

Un langage semble particulièrement adapté pour exprimer l'information selon les principes évoqués jusqu'ici : RDF (Resource Description Framework, voir (W3C, 2004). Il s'agit d'un dialecte XML développé par le W3C. Son but est la représentation de méta-données sous forme de graphes. À l'origine prévu pour des ressources Web, RDF peut néanmoins être utilisé avec n'importe quel type de données (méta données ou autres).

RDF doit être couplé à RDFS (RDF Schema) si l'on veut pouvoir introduire de la sémantique et ainsi respecter les contraintes d'intégrité dont nous avons déjà parlé. RDFS est un mécanisme qui permet de définir un vocabulaire particulier pour des données RDF et de spécifier les types d'objets sur lesquels les prédicats peuvent être appliqués. En pratique, RDFS nous permet de définir les relations, les entités et la manière dont ces éléments se combinent pour décrire l'information biographique.

Des langages tels que RQL (RDF Query Language) proposent une interrogation sous la forme d'une requête « select-from-where ». Celle-ci permet de spécifier un chemin, caractérisé par certaines contraintes, à parcourir dans les graphes RDF/RDFS (pattern-matching). Le résultat étant contenu dans la suite de nœuds du ou des chemins reconnus.

Des plateformes mettant en œuvre ces technologies existent déjà. L'un d'elles se nomme Sesame (Boekstra *et al.*, 2002 ; Aduna, 2005). Il s'agit d'une architecture « open source » (LGPL), indépendante du moyen d'enregistrement des données, qui permet le stockage persistent et l'interrogation de données RDF et RDFS. Sesame propose aussi, parmi d'autres choses, un module pour le dialogue avec un SGBD qui implémente la norme SQL92 et un langage d'interrogation étendant RQL, SeRQL.

6. Conclusion

L'extraction d'informations biographiques ne peut se passer d'une définition précise du domaine d'application. Cette définition peut être effectuée par la description, sous forme de relations entre entités, des évènements qui constituent la vie des personnes. Il s'agit là d'une tâche de longue haleine, qui n'a ici été qu'effleurée, mais qu'il est fondamental de continuer. La spécification ainsi obtenue sera extrêmement utile pour l'extraction d'informations par scénarios, la phase la plus complexe du processus. Les évènements biographiques déjà définis permettent également d'avancer des suggestions en ce qui concerne l'architecture logicielle à

mettre en œuvre pour le stockage des données. Une solution orientée vers la représentation de données sous forme de graphes doit permettre de stocker des triplets {sujet, relation, objet} dans une base de connaissances. Cette architecture reste cependant à évaluer et doit encore prouver sa faisabilité pratique.

Références

BOUHAFS A. (2004). « Système d'extraction d'information dédié à la veille. Qui est qui ? Qui fait quoi ? Où ? Quand ? Comment ? ». In *Actes de RECITAL 2004*.

BROEKSTRA J., KAMPMAN A., VAN HARMELEN F. (2002). « Sesame : A Generic Architecture for Storing and Querying RDF and RDF Schema ».In *Proceedings of the International Semantic Web Conference 2002*. Sardinia. http://www.openrdf.org/doc/papers/Sesame-ISWC2002.pdf.

CHARLET J., LAUBLET P., REYNAUD C. (2002). « Sur quelques aspects du Web sémantique ». In *Assises du GDR I3*. Nancy. http://www.lalic.paris4.sorbonne.fr/stic/articles/03-WebSemantique. pdf

CHINCHOR N. (1998). « MUC-7 Named Entity Task Definition (Version 3.5) ». In *Proceedings of MUC-7*, Fairfax.

FAIRON C., WATRIN P. (2003). « From extraction to indexation. Collecting new indexation keys by means of IE techniques ». In *Proceedings of EACL 2003*. Budapest.

FOUROUR N. (2004). *Identification et catégorisation automatiques des entités nommées dans les textes français*. Thèse de doctorat, Université de Nantes.

FRIBURGER N., MAUREL D. (2004). « Finite-state transducer cascades to extract named entities in texts ». In *Theoretical Computer Science* 313 (1) : 93-104.

GRISHMAN R. (1997). « Information extraction : Techniques and challenges ». In M.T. Pazienza (éd.), *Information Extraction : techniques and challenges*. Springer-Verlag, Berlin.

LE PRIOL F. (2001). « Identification, interprétation et représentation de relations sémantiques entre concepts ». In *Actes de TALN 2001*.

MALLCHOK F. (2004). *Automatic Recognition of Organisation Names in English Business News*. Thèse de doctorat, Université de Munich.

MCDONALD D.D. (1996). « Internal and External Evidence in the Identification and Semantic Categorization of Proper Names ». In B. Boguraev, J. Pustejovsky (éds), *Corpus processing for lexical acquisition* : 21-39.

PAUMIER S. (2004). *Unitex 1.2 Manuel d'utilisation*. Université de Marne-la-Vallée.

POIBEAU T., Nazarenko A. (1999). « L'extraction d'information, une nouvelle conception de la compréhension de texte ». In *TAL* 40 (2) : 87-115.

Références sur Internet

ADUNA B.V. (2005). « User Guide for Sesame (v1.2.3) ». http://www.openrdf.org/doc/sesame/users/.

DAVIS I., GALBRAITH D. (2004). « BIO: A vocabulary for biographical information ». http://purl.org/vocab/bio/.

KANZAKI (2003). « Who's who description vocabulary ». http://www.kanzaki.com/ns/whois.

W3C (2004). « Resource Description Framework (RDF) : Concepts and Abstract Syntax ». http://www.w3.org/ TR/2004/REC-rdf-concepts-20040210/.